

Health Care Predictive Analytics using Artificial Intelligence Techniques



Guanjin Wang

A thesis submitted for the Degree of
Doctor of Philosophy

Faculty of Engineering and Information Technology
University of Technology Sydney
July 2018

Certificate of Originality

This thesis is the result of a research candidature conducted jointly with another University as part of a collaborative Doctoral degree. This research is also supported by the Australian Government Research Training Program. I hereby declare that this thesis is my own work and that, to the best of my knowledge and belief, it reproduces no material previously published or written, nor material that has been accepted for the award of any other degree or diploma, except where due acknowledgment has been made in the text.

Student: Guanjin Wang

Date: 25/07/2018

Abstract

In recent years, advances in Artificial Intelligence (AI) are opening the door for intelligent health care data prediction and decision making. Machine learning, as an increasingly popular approach to AI, has been widely used to learn directly from data, adapt independently, and produce predictive outcomes, which support doctors when encountering complex health care predictive analytics. However, traditional machine learning methods are not always perfectly working in the health field, intrinsically due to little consideration for characteristic problems within health care data. For example, the small sample size problem is common due to complex data collection procedures and privacy concerns. Missing data is also widely encountered since most data are collected as a second-product of patient-care activities instead of following systematic research protocols. The class imbalance is another inevitable problem in the medical data as the normal class always predominates over the disease class. To solve aforementioned issues in health care predictive analytics, this study stands on the principles of machine learning and transfer learning to develop five advanced prediction models.

The first model is an output-based transfer least squares support vector machines (LS-SVMs) model which can leverage knowledge from the existing prediction model or on-line tool to facilitate the learning process on the current domain of interest with insufficient data. This model

overcomes the small sample size problem and improves the health care data prediction by learning knowledge from the other domain.

The second model is a novel additive LS-SVMs model which can make predictions simultaneously considering the influences on the classification error caused by missing features in a dataset. This model can generate valuable explanations regarding the influence levels of missing features for health professionals to improve the future data collection process.

The third model is a transfer-based additive LS-SVMs model which can deal with missing data from a transfer learning perspective. It can leverage the model knowledge learned from the complete portion of the dataset to help the learning process on the whole dataset with missing data. The proposed model can provide supplementary information for health professionals to improve the data quality via data cleaning.

The fourth model is a deep transfer additive LS-SVMs model called DTA-LS-SVMs and its imbalanced version called iDTA-LS-SVMs to enhance the prediction performance on the balanced and imbalanced datasets. Inspired by the stacked architecture and transfer learning mechanism, the model stacks multiple additive LS-SVMs based modules layer-by-layer and embeds model transfer between adjacent modules to guarantee their consistency.

The fifth model is a deep cross-output transfer LS-SVMs model called DCOT-LS-SVMs and its imbalanced version called IDCOT-LS-SVMs to improve the prediction performance on the balanced and imbalanced datasets. The cross-output transfer is used to transfer the predictive outcome from the previous module to the adjacent higher layer to achieve a better learning. Moreover, modules' parameters can be randomly assigned

in the proposed model which significantly reduces the time for model selection.

The proposed models are verified using experiments on the public UCI datasets. Moreover, case studies are conducted to validate and integrate the proposed models with real world applications, including bladder cancer prognosis, prostate cancer diagnosis, and predictions of elderly quality of life (QOL). The results have demonstrated that the models in this study can enhance the prediction performance while taking the characteristic problems within health care data into account, thus exhibiting promising potential for use in different health applications in future.

Acknowledgements

PhD study has been a challenging and memorial journey in the past three years. I would like to extend my warmest gratitude to the people who inspired and helped me along this journey.

First, I would like to express my earnest thanks to my chief supervisor in the Hong Kong Polytechnic University, Professor Choi Kup Sze, and my chief supervisor in the University of Technology Sydney, Distinguished Professor Jie Lu for their continuous support, precious guidance and enormous patience throughout my study. Without their excellent supervision and valuable suggestions, this Joint-PhD study between two universities could not have been finished. Thank you for all your detailed comments and suggestions on my research, manuscripts and presentations. Your strict academic attitude and hard working style have deeply influenced me and will always inspire and motivate me in my future work and life. I also would like to address my sincere thanks to Professor Guangquan Zhang for his academic suggestions and advices, Dr Chiang Chung Lim Vico for his health care knowledge sharing and advices and Dr Lam Kin Man for his medical knowledge sharing and data support.

I am honored to have met all the great researchers and staffs in the Centre for Smart Health in the Hong Kong Polytechnic University and Centre for Artificial Intelligence in the University of Technology Sydney. I appreciate all their valuable and critical comments during my presentations, and the discussions with them were enlightening.

I am grateful to the School of Nursing in the Hong Kong Polytechnic University and the School of Software in the Faculty of Engineering and Information Technology in the University of Technology Sydney. This study was supported by the YC Yu Scholarship for the Centre for Smart Health, UTS Doctoral Scholarship, the Research Grants Council of the Hong Kong SAR and the Australian Research Council (ARC) discovery project.

Lastly, I would like to express my most grateful appreciation to my family for their unconditional love, encouragement and support. This journey would not have been possible without their help. I am especially grateful to my parents. Thank you for being the first teacher in my life and guiding me to be a better person. I always know you believe in me more than I do in myself.

Contents

Contents	vii
List of Figures	xii
List of Tables	xiv
1 Introduction	1
1.1 Health Care Data Analytics	1
1.2 Challenges	3
1.3 Research Contributions	4
1.4 Research Significance	6
1.5 Thesis Structure	6
2 Literature Review	9
2.1 Health Care Data Prediction: Classical Statistical Models	9
2.2 Health Care Data Prediction: Artificial Intelligence Models	12
2.2.1 Artificial Neural Networks	13
2.2.2 Support Vector Machines	14
2.2.3 Least Square Support Vector Machines	16
2.2.4 Naive Bayes Classifiers	17
2.2.5 Extreme Learning Machines	17

2.2.6	k -nearest Neighbors Algorithms	18
2.3	Transfer Learning	19
2.3.1	Definition and Notations	21
2.3.2	A Categorization of Transfer Learning Techniques	22
2.4	Missing Data Problem and Solutions	25
2.5	Class Imbalance Problem and Solutions	28
2.6	Deep and Shallow Architectures	30
3	An Output-based Transfer LS-SVMs Model for Bladder Cancer Prognosis	
	with Insufficient Data	33
3.1	Introduction	33
3.2	Output-based Transfer LS-SVMs Model with Insufficient Data	34
3.2.1	Inverted Pyramid Dataset	35
3.2.2	Framework of the Proposed Model	35
3.2.3	Handle Probabilistic Outputs From the Existing Model	38
3.2.4	Output-based Transfer LS-SVMs in Target Domain	38
3.2.5	Fast Leave-one-out Cross Validation Strategy for Parameter Tuning	44
3.2.6	Computational Complexity	48
3.3	A Case Study on a Real World Bladder Cancer Dataset	49
3.3.1	Data Collection and Existing Prediction Model	49
3.3.2	Experimental Design	50
3.3.3	Results Analysis	54
3.4	Summary	55
4	A Novel Additive LS-SVMs Model for Predicting Elderly QOL with	
	Missing Data	58
4.1	Introduction	58

4.2	Novel Additive LS-SVMs Model with Missing Data	59
4.2.1	Problem Description	59
4.2.2	Novel Additive LS-SVMs Model	60
4.2.3	Fast Leave-one-out Cross Validation Strategy	63
4.2.3.1	Fast Leave-one-out Cross Validation for Parameter Tuning	63
4.2.3.2	Interpretation of Influences of Missing Features . .	64
4.3	A Case Study on a Real World Community Health Care Dataset . . .	65
4.3.1	Data Collection	65
4.3.2	Data pre-processing	69
4.3.3	Results Analysis	69
4.4	Summary	72
5	A Transfer-based Additive LS-SVMs Model for Predicting Elderly QOL with Missing Data	74
5.1	Introduction	74
5.2	Transfer-based Additive LS-SVMs Model with Missing data	76
5.2.1	Problem Description	76
5.2.2	Framework of the Proposed Model	77
5.2.3	Adaptive Regularization	77
5.2.4	Transfer-based Additive LS-SVMs Model	79
5.2.5	Fast Leave-one-out Cross Validation Strategy	82
5.2.5.1	Fast Leave-one-out Cross Validation for Parameter Tuning	82
5.2.5.2	Interpretation of Influences of Incomplete Samples .	83
5.2.6	Computational Complexity	84
5.3	Experiments	85
5.3.1	UCI Datasets	85

5.3.2	Experimental Design	86
5.3.3	Experimental Results Analysis	87
5.4	A Case Study on a Real World Community Health Care Dataset . . .	89
5.4.1	Data Collection and Pre-processing	89
5.4.2	Challenge	89
5.4.3	Results Analysis	90
5.4.4	Contribution	91
5.5	Summary	99
6	A Deep Transfer Additive LS-SVMs Model for Predicting Elderly QOL with Imbalance Data	103
6.1	Introduction	103
6.2	Deep Transfer Additive LS-SVMs Model	105
6.2.1	Framework of the Proposed Model	105
6.2.2	Deep Transfer Additive LS-SVMs Model	107
6.2.3	Fast Leave-one-out Cross Validation Strategy	110
6.2.4	Computational Complexity	111
6.3	Extension on Class Imbalance Problems	114
6.4	Experiments	115
6.4.1	UCI datasets	116
6.4.2	Parameter Setup	117
6.4.3	Experimental Results Analysis	117
6.5	A Case Study on a Real World Community Health Care Dataset . . .	125
6.5.1	Data Collection	125
6.5.2	Experimental Design	125
6.5.3	Results Analysis	126
6.6	Statistical Analysis	129
6.7	Summary	131

7	A Deep Cross-output Transfer LS-SVMs Model for Diagnosing Prostate Cancer with Imbalance Data	133
7.1	Introduction	133
7.2	Deep Cross-output Transfer LS-SVMs Model	135
7.2.1	Framework of the Proposed Model	135
7.2.2	Cross-output Knowledge Transfer Under a Stacked Architecture	137
7.2.3	Fast Leave-one-out Cross Validation Strategy	140
7.2.4	Computational Complexity	141
7.3	Extension on Class Imbalance Problems	144
7.4	Experiments	145
7.4.1	UCI Datasets	145
7.4.2	Parameter Setup	146
7.4.3	Experimental Results Analysis	146
7.5	A Case Study on a Real World Prostate Cancer Dataset	151
7.5.1	Data Collection	151
7.5.2	Results Analysis	152
7.5.3	Contribution	153
7.6	Statistical Analysis	153
7.7	Summary	157
8	Conclusion and Future Work	158
8.1	Conclusions	158
8.2	Future Study	160
9	Publications during PhD Study	162
	Bibliography	164

List of Figures

1.1	THESIS STRUCTURE	8
2.1	THE STRUCTURE OF A BIOLOGICAL NEURON	14
2.2	SVM LEARNS THE HYPERPLANE THAT BEST SEPARATES THE TWO CLASSES WITH DATA POINTS DENOTED BY CIRCLES AND TRIAN- GLES, WITH THE LABEL AND RESPECTIVELY.	15
2.3	EXAMPLE OF k -NN CLASSIFICATION	19
2.4	DIFFERENT LEARNING PROCESSES OF TRADITIONAL MACHINE LEARNING AND TRANSFER LEARNING	20
2.5	TWO WAYS TO EXPAND THE FEATURE SPACE IN DEEP STACKED ARCHITECTURE	32
3.1	THE INVERTED PYRAMID DATASET IN WHICH $d' < d$	36
3.2	THE FRAMEWORK OF THE PROPOSED MODEL	37
3.3	ONLINE NOMOGRAM PREDICTING THE PROBABILITY OF MORTAL- ITY DUE TO BLADDER CANCER VERSUS OTHER CAUSES	53
3.4	ROC CURVE OF THE PROPOSED MODEL-V1 AND COMPARATIVE METHODS	56
3.5	ROC CURVE OF THE PROPOSED MODEL-V2 AND COMPARATIVE METHODS	57

LIST OF FIGURES

4.1	PATTERN CLASSIFICATION ON (A) COMPLETE AND (B) INCOMPLETE DATASET	60
5.1	DATASET REPRESENTATION	76
5.2	THE FRAMEWORK OF THE PROPOSED TRANSFER-BASED ADDITIVE LS-SVMs	77
5.3	COMPARATIVE RESULTS FOR THE COMMUNITY HEALTH CARE DATASET	100
5.4	COMPARATIVE RESULTS AFTER DATA CLEANING FOR THE COMMUNITY HEALTH CARE DATASET	100
5.5	COMPARATIVE RESULTS OF PROPOSED AND COMPARATIVE METHODS ON SEVEN PUBLIC DATASETS	101
6.1	THE HIERARCHICAL ARCHITECTURE AND LEARNING PROCESS OF DTA-LS-SVMs MODEL	106
6.2	THE AUGMENTED SPACE \vec{X}'_l OF \vec{X}	107
7.1	THE STACKED ARCHITECTURE AND LEARNING PROCESS IN DCOT-LS-SVMs	136

List of Tables

3.1	BASILINE CHARACTERISTICS OF THE COHORT	51
3.2	THE INPUTS AND OUTPUT OF THE PREDICTION MODEL	52
3.3	PARAMETER SETTINGS OF THE PROPOSED AND COMPARATIVE METHODS	54
3.4	PERFORMANCE RESULTS OF THE PROPOSED MODELS AND COM- PARATIVE METHODS	56
4.1	THE EXTENT OF MISSING DATA IN CERTAIN FEATURES ($N = 444$) .	67
4.2	HEALTH RELATED ASSESSMENTS AND QUESTIONNAIRE ON MIHC	68
4.3	RE-CATEGORIZATION OF THE RESPONSES TO OVERALL QOL . . .	69
4.4	CLASSIFICATION ACCURACIES OF THE PROPOSED AND COMPARA- TIVE METHODS	70
4.5	INFLUENCES OF MISSING FEATURES	71
5.1	DATASET DESCRIPTIONS	85
5.2	PERFORMANCE RESULTS FOR THE <i>Surgery</i> DATASET	92
5.3	PERFORMANCE RESULTS FOR THE <i>Diabetic</i> DATASET	93
5.4	PERFORMANCE RESULTS FOR THE <i>Pima</i> DATASET	94
5.5	PERFORMANCE RESULTS FOR THE <i>Bupa</i> DATASET	95
5.6	PERFORMANCE RESULTS FOR THE <i>Breast</i> DATASET	96
5.7	PERFORMANCE RESULTS FOR THE <i>Titanic</i> DATASET	97

LIST OF TABLES

5.8	PERFORMANCE RESULTS FOR THE <i>German</i> DATASET	98
5.9	PERFORMANCE RESULTS FOR THE COMMUNITY HEALTH CARE DATASET	98
5.10	PERFORMANCE RESULTS AFTER DATA CLEANING FOR THE COM- MUNITY HEALTH CARE DATASET	99
5.11	AVERAGE RANKINGS OF THE PROPOSED AND COMPARATIVE METH- ODS ON SEVEN PUBLIC DATASETS IN TERMS OF AVERAGE ACCU- RACY (p -VALUE=0.000704)	99
5.12	HOLM POST-HOC COMPARISON RESULTS FOR THE PROPOSED AND COMPARATIVE METHODS IN TERMS OF AVERAGE ACCURACY WITH $\alpha = 0.05$	102
6.1	UCI DATASETS DESCRIPTION	116
6.2	PERFORMANCE RESULTS ON BALANCED UCI DATASETS	118
6.3	PERFORMANCE RESULTS ON IMBALANCED UCI DATASETS	118
6.4	PERFORMANCE RESULTS ON THE <i>Australian</i> DATASET WITH DIF- FERENT RATIOS OF TRAINING AND TESTING DATA	119
6.5	PERFORMANCE RESULTS ON THE <i>Diabetic</i> DATASET WITH DIFFER- ENT RATIOS OF TRAINING AND TESTING DATA	119
6.6	PERFORMANCE RESULTS ON THE <i>credit approval</i> DATASET WITH DIFFERENT RATIOS OF TRAINING AND TESTING DATA	119
6.7	PERFORMANCE RESULTS ON THE <i>mammographic</i> DATASET WITH DIFFERENT RATIOS OF TRAINING AND TESTING DATA	120
6.8	PERFORMANCE RESULTS ON THE <i>breast cancer</i> DATASET WITH DIFFERENT RATIOS OF TRAINING AND TESTING DATA	120
6.9	PERFORMANCE RESULTS ON THE <i>Pima Indians</i> DATASET WITH DIFFERENT RATIOS OF TRAINING AND TESTING DATA	120

LIST OF TABLES

6.10	PERFORMANCE RESULTS ON THE <i>Indians liver</i> DATASET WITH DIFFERENT RATIOS OF TRAINING AND TESTING DATA	121
6.11	TRAINING AND TESTING TIME (SECONDS) ON SEVEN UCI DATASETS	124
6.12	PERFORMANCE RESULTS ON THE COMMUNITY HEALTH CARE DATASET USING iDTA-LS-SVMs	128
6.13	PERFORMANCE RESULTS ON THE COMMUNITY HEALTH CARE DATASET USING DTA-LS-SVMs AND THE OTHER COMPARATIVE METHODS	128
6.14	TRAINING AND TESTING TIME (SECONDS) ON THE COMMUNITY HEALTH CARE DATASET	128
6.15	AVERAGE RANKINGS OF DTA-LS-SVMs AND THE COMPARATIVE METHODS ON BALANCED DATASETS IN TERMS OF ACCURACY (p - VALUE= 0.049787)	130
6.16	HOLM POST-HOC COMPARISON RESULTS FOR DTA-LS-SVMs AND THE OTHER METHODS IN TERMS OF ACCURACY WITH $\alpha = 0.05$	130
6.17	AVERAGE RANKINGS OF DTA-LS-SVMs AND THE COMPARATIVE METHODS ON BALANCED DATASETS IN TERMS OF F1-SCORE (p - VALUE= 0.038774)	130
6.18	HOLM POST-HOC COMPARISON RESULTS FOR DTA-LS-SVMs AND THE OTHER METHODS IN TERMS OF F1-SCORE WITH $\alpha = 0.05$	131
6.19	AVERAGE RANKINGS OF iDTA-LS-SVMs AND THE COMPARA- TIVE METHODS ON IMBALANCED DATASETS IN TERMS OF F1- SCORE (p -VALUE= 0.022371)	131
6.20	HOLM POST-HOC COMPARISON RESULTS FOR iDTA-LS-SVMs AND THE OTHER METHODS WITH $\alpha = 0.05$	131
7.1	UCI DATASETS DESCRIPTION	146
7.2	PERFORMANCE RESULTS ON BALANCED DATASETS	148

LIST OF TABLES

7.3	PERFORMANCE RESULTS ON IMBALANCED UCI DATASETS	148
7.4	PERFORMANCE RESULTS ON THE <i>Aus</i> DATASET	148
7.5	PERFORMANCE RESULTS ON THE <i>Diabetic</i> DATASET	149
7.6	PERFORMANCE RESULTS ON THE <i>Credit</i> DATASET	149
7.7	PERFORMANCE RESULTS ON THE <i>Mammographic</i> DATASET	149
7.8	PERFORMANCE RESULTS ON THE <i>Breast</i> DATASET	150
7.9	PERFORMANCE RESULTS ON THE <i>Pima</i> DATASET	150
7.10	PERFORMANCE RESULTS ON THE <i>ILPD</i> DATASET	150
7.11	TRAINING AND TESTING TIME (SECONDS) ON SEVEN UCI DATASETS	151
7.12	BASLINE CHARACTERISTICS OF THE COHORT	154
7.13	PERFORMANCE RESULTS ON THE PROSTATE CANCER DATASET . .	154
7.14	TRAINING AND TESTING TIME (SECONDS) ON THE PROSTATE CANCER DATASET	154
7.15	AVERAGE RANKINGS OF DCOT-LS-SVMs AND THE COMPARA- TIVE METHODS IN TERMS OF ACCURACY ($p=0.022371$)	155
7.16	HOLM POST-HOC COMPARISON RESULTS FOR DCOT-LS-SVM AND THE OTHER METHODS IN TERMS OF ACCURACY WITH $\alpha = 0.05$	156
7.17	AVERAGE RANKINGS OF DCOT-LS-SVMs AND THE COMPARA- TIVE METHODS IN TERMS OF AUC ($p = 0.022371$)	156
7.18	HOLM POST-HOC COMPARISON RESULTS FOR DCOT-LS-SVMs AND THE OTHER METHODS IN TERMS OF AUC WITH $\alpha = 0.05$. .	156
7.19	AVERAGE RANKINGS OF IDCOT-LS-SVMs AND THE COMPARA- TIVE METHODS IN TERMS OF AUC ($p = 0.038774$)	156
7.20	HOLM POST HOC COMPARISON RESULTS FOR IDCOT-LS-SVMs AND THE OTHER METHODS IN TERMS OF AUC WITH $\alpha = 0.05$. .	156